Proceedings of the 1st International Scientific and Practical Conference «Modern Knowledge: Research and Discoveries»

(May 19-20, 2023). Vancouver, Canada

No 155

INFORMATION AND WEB TECHNOLOGIES

DOI 10.51582/interconf.19-20.05.2023.039

Vision transformer for skin cancer classification

Nikitin Vladyslav[1],

Shapoval Nataliia[2]

[1] MSc in Computer Science;
Institute for Applied System Analysis, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»; Ukraine

[2] Candidate of Engineering Sciences;
Institute for Applied System Analysis, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»; Ukraine

Abstract.
This paper investigates the use of vision transformers (ViT) for skin cancer classification tasks, compared to convolutional models. We propose a novel ViT architecture that effectively classifies skin cancer images. Our findings suggest that ViT models have the potential to outperform convolutional models, especially with larger datasets.

Keywords:
skin cancer
machine learning
computer vision
vision transformer
attention
convolutional neural networks

Proceedings of the 1st International Scientific and Practical Conference «Modern Knowledge: Research and Discoveries»

(May 19-20, 2023). Vancouver, Canada

No 155

INFORMATION AND WEB TECHNOLOGIES

INTRODUCTION

Skin cancer is an uncontrolled growth of pathological cells in the epidermis caused by repaired DNA damage that triggers mutations. Mutations force skin cells to multiply rapidly and create harmful tumours. The main types of skin cancer are basal cell carcinoma, squamous cell carcinoma, and melanoma. All types of skin cancer can be caused by two main factors. First, harmful solar ultraviolet radiation and ultraviolet solariums . And second, hereditary genetic predisposition to this disease [1].

When this disease is diagnosed at the early stage, complete removal of the tumour with minimal scarring or no scarring at all is possible with a high probability, not to mention the minimal risk to health. This is especially true for cases where the doctor detects a tumour before it transforms into a malignant one and penetrates under the skin layer [1].

On the other hand, late diagnosis is very likely to determine a negative outcome. For example, for melanoma, the 5-year mortality rate can range from 30 to 68% depending on the location of the tumour [2]. In Ukraine, we have slightly different but still sad statistics. As of 2018, 2835 cases of melanoma were detected. 10.9% of these patients did not survive one year from the moment of diagnosis [3].

The automation of the skin cancer detection process could significantly increase the percentage of patients who were detected at an early stage and accordingly increase the chances of successful treatment of the disease. Therefore, the purpose of this work is to study the possibility of early detection of skin cancer (in particular, melanoma) using visual transformers - machine learning models derived from the transformer model applied in natural language processing. This new model outperforms standard approaches in many tasks, such as various types of convolutional neural networks.

VISION TRANSFORMER AND ATTENTION

The article on the vision transformer (ViT) "An Image is Worth 16x16 Words" [4] demonstrates the implementation of a pure transformer model without the need for convolutional layers. The article shows how using a vision transformer can yield better results than using any convolutional model in image recognition tasks, while using relatively fewer

INFORMATION AND WEB TECHNOLOGIES

resources [4].

As can be seen on Fig. 1, the image is getting split to patches, flattened and passed to attention func. After getting the attention map, the needed analysis can be performed.

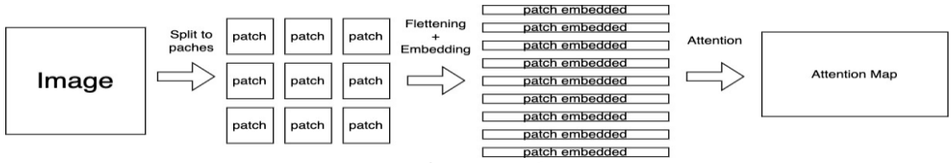

Figure 1

ViT scheme

The mechanism of attention was introduced in 2014 as a solution to the problem that arises when using a fixed-length encoding vector in encoder-decoder type models. The issue is that the decoder may have limited access to the information presented in the input model.

In general, when computing the attention function, three key components are used: Query, Key, and Value. These can be three identical matrices. The attention value is computed using matrix multiplication of the Query and Key, divided by the square root of the dimension of the Key, followed by the application of the softmax function and matrix multiplication with the Value (Fig. 1) [5].

In this work, the multi-head attention function was used, which applies multiple attention functions to the patches simultaneously to extract different features.

DATASET AND PREPROCESSING

To train the model, a dataset synthesised from the datasets PH2 [6], MED-NODE [7], HAM10000 [8], ISIC [9] was used (Fig. 2). Mentioned dataset were selectively taken to create a balanced skin cancer dataset with maximum number of images.

For image preprocessing, images from the dataset were scaled to 100x100. Options for increasing image contrast and converting to a single channel (black and white) were also considered. However, training the models on images with original colours proved to be more effective. To increase the size of the dataset and thus improve the quality of the model, augmentation was used, including horizontal flipping and scaling of the images.

INFORMATION AND WEB TECHNOLOGIES

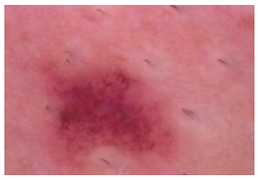

Figure 2
Example of an image from the training dataset

ARCHITECTURES

Three variants of the ViT architecture were considered in this work. The base architecture of the model is presented in Figure 4. It includes an input layer, patching, patch encoding, 1-10 layers of transformers, and classification. The model takes as input a set of three-channel 100x100 images. Next comes patching. The number and size of patches are parameters of the model. For example, suppose we keep the image size at 100x100 and have patches of size 10. Then each patch is unfolded into a sequence of pixels (all three channels together), so the resulting image size becomes 300x100, that is, 100 patches and 300 pixels in one patch (three channels). During patch encoding, position embeddings are added to the patches. In tables, this architecture will be referred to as *base_vit* (Fig. 3).

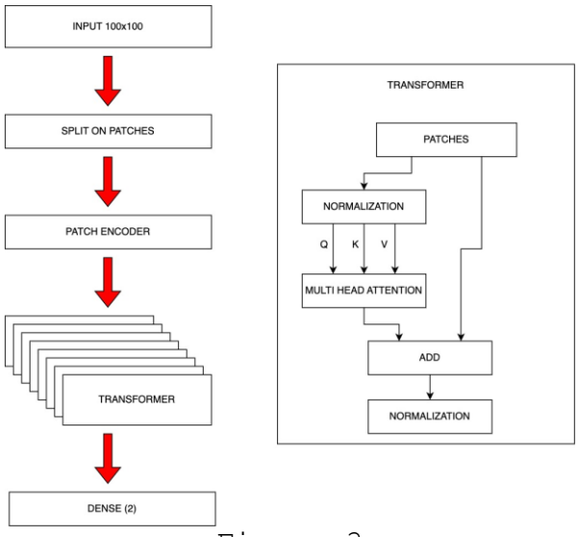

Figure 3
Architecture of base_vit

INFORMATION AND WEB TECHNOLOGIES

The second proposed approach is using CrossViT [10] (Fig. 4). It can be particularly useful for focusing attention on specific areas of an image, which is especially important when dealing with skin lesions that may not always fit neatly within patches. The architecture introduced here will be referred to as *cross_vit* in tables, and it represents an innovative approach to skin cancer classification using vision transformer architectures.

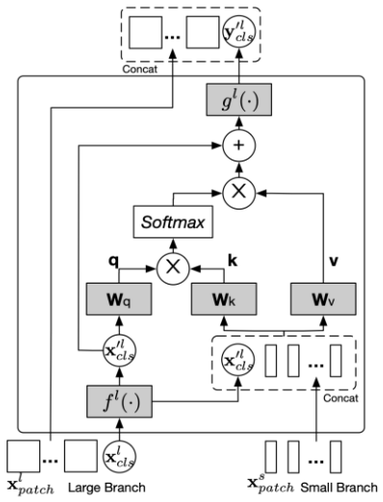

Figure 4

CrossViT, attention scheme [10]

The skin lesions may not always be evenly distributed across patches. As a result, the edges may not always fall within the attention areas. An example can be seen in Figure 5.

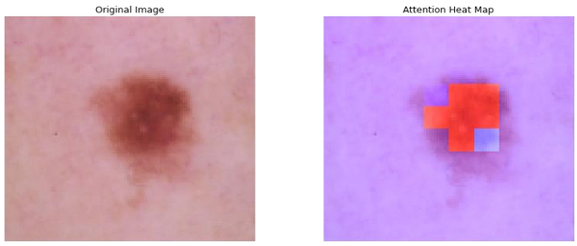

Figure 5

Heatmap of attention for *base_vit*

INFORMATION AND WEB TECHNOLOGIES

To address this issue, the use of additional larger patches was proposed to increase the attention value of the surrounding patches (Fig. 6). However, the attention function works equally in parallel for patches of both sizes at each application of the attention function. Before the next layer of the attention function, the attention values of the larger patches are projected onto the attention values of the smaller patches, added, and normalised (Fig. 7). As a result, the attention value for the surrounding patches increases relative to the centre. This architecture will be referred to as *merge_vit* in the work.

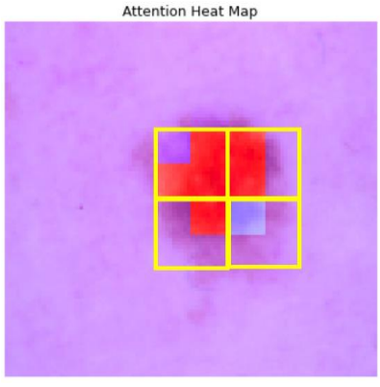

Figure 6

The values of the yellow patches will be added to the smaller patches they touch. As a result, patches on the periphery can have a higher attention value

The basic parameters of the model are *img_size*, *patch_size*, *tfrm_layers*, *proj_dim*, and *att_heads*. *img_size* corresponds to the size to which the image will be reduced during training. patch_size corresponds to the size of one side of the square patch to which the image will be divided. *tfrm_layers* is how many times the multi-attention function will be applied to the image. *proj_dim* is the dimension to which the patch will be reduced. *att_heads* corresponds to how many single attention functions will be applied to the image.

For models that have two types of patches, the *patch_size* parameter will be specified in the format (x, y), where x and y are the sizes of the smaller and larger patch types, respectively.

INFORMATION AND WEB TECHNOLOGIES

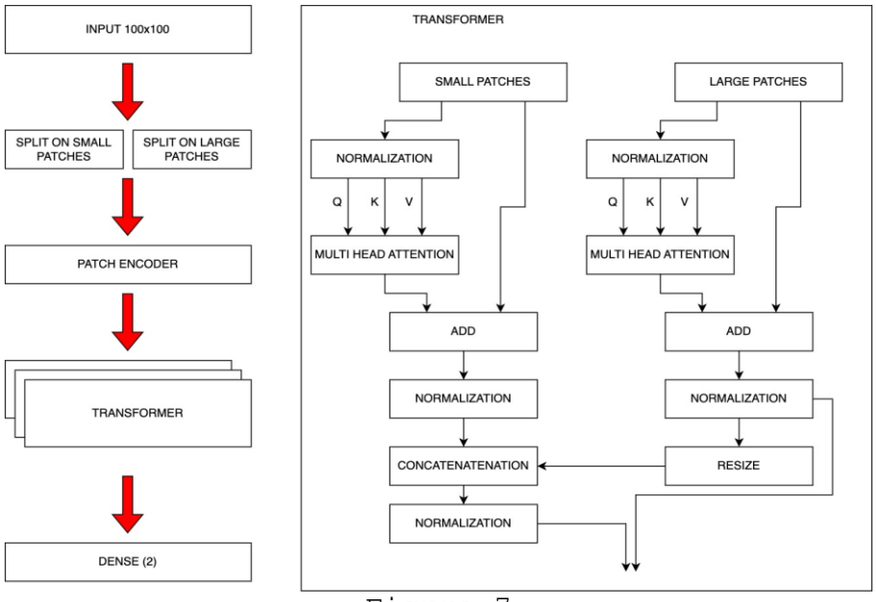

Figure 7

The figure shows a diagram of a vision transformer using larger patches. After each attention iteration, the values of the larger patches are added to the smaller ones

METRICS

The main metric used to monitor the model during training and determine the best model is accuracy, which is the percentage of images from the test set that are classified correctly.

In addition, errors of the first and second kind were taken into account for comparison. Since the task is medical, the accepted standard is that a "first'' kind error is better than a second kind error. In our case, this is also true because a misclassified nevus as cancer may only have the worst consequences as an additional visit to the doctor. On the other hand, a second kind error, where melanoma is classified as a nevus, can lead to a delay in the start of treatment, which is the worst possible consequence of an error.

It is worth noting that comparing the absolute values of errors of the first and second kind in this study makes sense because the division of the sample into training and testing was done once, independently of the training of any of the

INFORMATION AND WEB TECHNOLOGIES

models, so all models have the same training and test sets.

Other traditional classification metrics such as precision, recall, f1-score, support, and training charts will only be provided for the most effective models.

LEARNING

AdamW optimizer from the tensorflow-addon framework was used. This is a stochastic optimizer for machine learning models that modifies the standard weight decay in the Adam algorithm, making it less dependent on the updated gradient value.

The table shows the results of training vision transformers with different architectures and parameters. Ideally, the same parameters should be selected for different architectures, but hardware powers were limited. Therefore, the model parameters were chosen to be as close to each other as possible.

The left side of the table shows the parameters with which the model and its architecture were created. On the right side, there are three key metrics that were used to select the most efficient models. In the future, when referring to any of the models in the table, the notation architecture_size_patch will be used.

As can be seen from Table 1, the most efficient models were base_vit_12 and merge_vit_12.

Table 1

Comparative table of models based on vision transformers

Parameters						**Metrics**		
							errors	
Architecture	*img_size*	*patch_size*	*tfrm_layers*	*proj_dim*	*att_heads*	*accuracy, %*	*Type 1*	*Type 2*
base_vit	100	10	8	64	4	76.05	35	45
base_vit	72	6	8	64	4	74.55	52	33
base_vit	36	2	6	36	4	72.75	50	41
base_vit	12	1	6	12	3	78.14	45	28
cross_vit	100	(10, 20)	8	64	3	67.22	70	39
cross_vit	72	(6, 8)	6	64	3	68.34	98	7
cross_vit	36	(2, 6)	6	36	2	74.12	81	5
cross_vit	12	(1, 4)	6	12	2	72.11	26	67
merge_vit	100	(10, 20)	8	64	3	60.48	132	0

INFORMATION AND WEB TECHNOLOGIES

Table continuation 1

merge_vit	72	(6, 12)	6	48	3	70.66	74	24
merge_vit	24	(2, 6)	6	24	2	73.65	61	27
merge_vit	12	(1, 3)	8	12	3	81.14	33	30

RESULTS

Below are the extended metrics of the models that showed the best results during training (Fig. 8, Fig. 9, Fig. 10, Fig. 11, Table 2, Table 3).

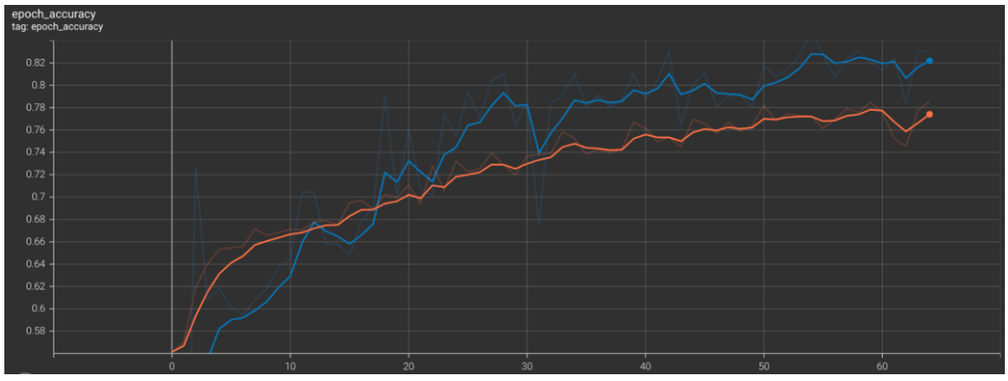

Figure 8

Learning accuracy of the model *base_vit_12*

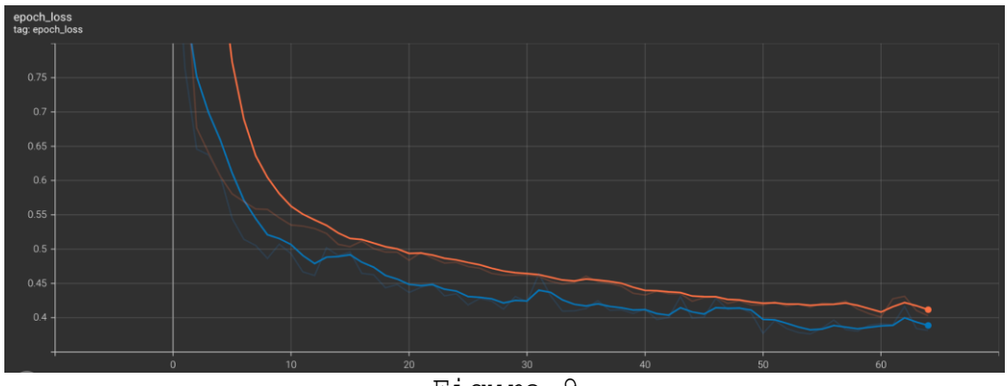

Figure 9

learning loss of the model *base_vit_12*

Table 2

Classification metrics of the model *base_vit_12*

	precision	**recall**	**f1-score**	**samples**
Not Cancer	0.79	0.86	0.83	202
Cancer	0.76	0.66	0.70	132

INFORMATION AND WEB TECHNOLOGIES

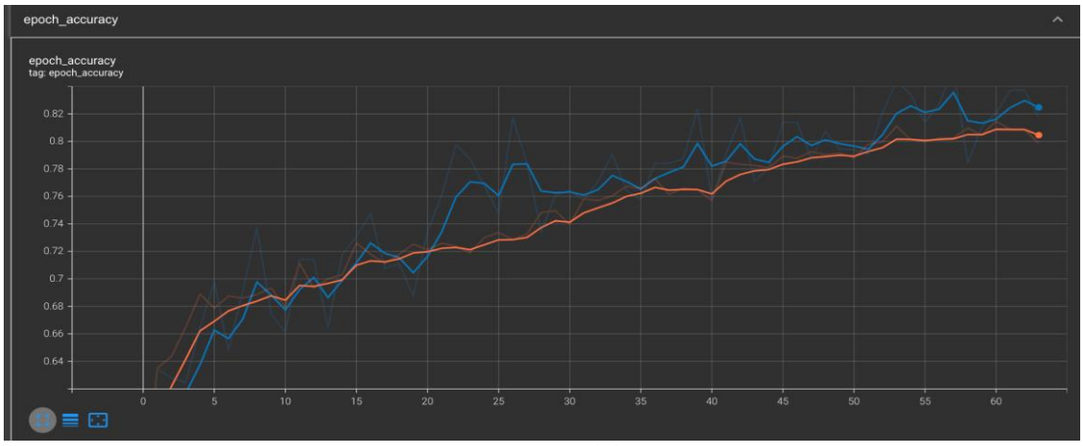

Figure 10

Learning accuracy of the model merge_vit_12

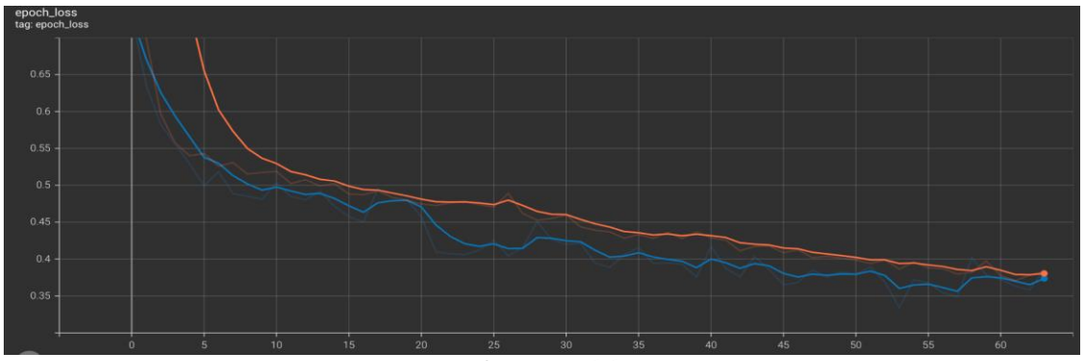

Figure 11

Learning loss of the model merge_vit_12

Table 3

Classification metrics of the model loss *merge_vit_12*

	precision	**recall**	**f1-score**	**samples**
Not Cancer	0.84	0.85	0.85	202
Cancer	0.77	0.75	0.76	132

COMPARISON WITH CNN

For comparison, the architecture from an existing study on skin cancer classification [11] was taken. The authors of the vision transformer claim that it outperforms convolutional models on large datasets. To compare, training on subsets of the overall training dataset with sizes of 100, 200, 500, 1500 (the whole dataset is 3005) is proposed to

evaluate the dependence of key metrics of each model on the dataset size.

Table 4

Comparison of convolution model and ViT on different dataset sizes

parameters		**Metrics**		
Model	*Dataset size*	*accuracy, %*	*Errors*	
			Type 1	*Type 2*
cnn	100	60.48	132	0
cnn	200	60.48	132	0
cnn	500	45.51	80	102
cnn	1500	64.37	88	31
cnn	3005	68.26	90	16
base_vit_12	100	60.48	132	0
base_vit_12	200	60.48	132	0
base_vit_12	500	60.48	132	0
base_vit_12	1500	68.34	73	33
base_vit_12	3005	78.14	45	28
merge_vit_12	100	60.48	132	0
merge_vit_12	200	60.48	132	0
merge_vit_12	500	68.34	93	13
merge_vit_12	1500	76.35	32	47
merge_vit_12	3005	81.14	33	30

As shown in Table 4, the cnn model performed better than the base_vit_12 model for small datasets, but merge_vit_12 showed the best results, particularly as the dataset size increased. In conclusion, it can be inferred that the quality of vision transformers improves faster with dataset size, which means that transformer models may potentially have a considerable advantage over convolutional models in terms of efficiency for large datasets.

CONCLUSION

In conclusion, skin cancer is a serious and prevalent disease that can be fatal if not diagnosed and treated early. In recent years, there has been a growing interest in the use of deep learning models for skin cancer classification, and the ViT architecture has shown great promise in this field.

We have discussed the ViT architecture and its potential for skin cancer classification. Specifically, we have compared the performance of three different ViT models: basic ViT, ViT with Cross Attention, and a modified version of basic ViT. Our experiments have shown that ViT outperforms basic

INFORMATION AND WEB TECHNOLOGIES

CNN models in terms of accuracy, and that the accuracy of ViT models grows more quickly with increasing dataset sizes.

Based on these results, we can conclude that ViT has significant potential for improving the accuracy of skin cancer classification. While more research is needed to fully explore the capabilities of this architecture, our findings suggest that ViT should be considered as a promising tool for this important task. By continuing to investigate and develop ViT models for skin cancer classification, we can help to improve the accuracy of diagnosis and potentially save lives.

References:

[1] *Skin cancer information* (2023) *The Skin Cancer Foundation*. Available at: https://www.skincancer.org/skin-cancer-information/ (Accessed: May 1, 2023).
[2] *Skin cancer* (no date) *American Academy of Dermatology*. Available at: https://www.aad.org/media/stats-skin-cancer (Accessed: May 1, 2023).
[3] Adjusted rates 2018 melanoma of skin C43 table 1 - general rates, 2018 (no date). Available at: http://www.ncru.inf.ua/publications/BULL_21/PDF_E/38-39-mel.pdf (Accessed: May 1, 2023).
[4] Dosovitskiy, A. *et al.* (2021) *An image is worth 16x16 words: Transformers for image recognition at scale*, *arXiv.org*. Available at: https://arxiv.org/abs/2010.11929 (Accessed: May 1, 2023).
[5] *Vaswani, A. et al. (2017) Attention is all you need, arXiv.org. Available at: https://arxiv.org/abs/1706.03762 (Accessed: May 1, 2023).*
[6] *PH² Database* (no date) *Addi - automatic computer-based diagnosis system for dermoscopy images*. Available at: https://www.fc.up.pt/addi/ph2%20database.html (Accessed: May 1, 2023).
[7] *MED-NODE* (no date) *Dermatology database used in Med-node*. Available at: https://www.cs.rug.nl/~imaging/databases/melanoma_naevi/ (Accessed: May 1, 2023).
[8] Tschandl, P. (2023) *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*, *Harvard Dataverse*. Harvard Dataverse. Available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FDBW86T (Accessed: May 1, 2023).
[9] *Isic Archive* (no date) *ISIC Archive*. Available at: https://www.isic-archive.com/ (Accessed: May 1, 2023).
[10] Chen, C.-F., Fan, Q. and Panda, R. (2021) *Crossvit: Cross-attention multi-scale vision transformer for Image Classification*, *arXiv.org*. Available at: https://arxiv.org/abs/2103.14899 (Accessed: May 1, 2023).
[11] Tirth Patel (no date) *Tirth27/Skin-cancer-classification-using-deep-learning*, *GitHub*. Available at: https://github.com/Tirth27/Skin-Cancer-Classification-using-Deep-Learning (Accessed: May 1, 2023).